基于LSTM模型的中文图书多标签分类研究*

邓三鸿　傅余洋子　王　昊
(南京大学信息管理学院　南京 210023)
(江苏省数据工程与知识服务重点实验室(南京大学)　南京 210023)

摘要:【**目的**】利用LSTM模型和字嵌入的方法构建分类系统, 提出一种中文图书分类中多标签分类的解决方案。【**方法**】引入深度学习算法, 利用字嵌入方法和LSTM模型构建分类系统, 对题名、主题词等字段组成的字符串进行学习以训练模型, 并采用构建多个二元分类器的方法解决多标签分类问题, 选择3所高校5个类别的书目数据进行实验。【**结果**】从整体准确率、各类别精度、召回率、F1值多个指标进行分析, 本文提出的模型均有良好表现, 有较强的实际应用价值。【**局限**】数据仅涉及中图分类法5个类别, 考虑的分类粒度较粗等。【**结论**】基于LSTM模型的中文图书分类系统具有预处理简单、增量学习、可迁移性高等优点, 具备可行性和实用性。

关键词: LSTM模型　深度学习　字嵌入　图书自动分类　多标签分类
分类号: TP391

1 引　言

近年来, 信息技术飞速发展, 人类已经进入了大数据时代。而这一变化也蔓延到图书情报领域, 其典型现象之一就是数字图书馆的发展。数字图书馆利用现代化的数字信息技术,特别是互联网技术, 延伸传统图书馆的职能, 从而更好地组织和传递文献信息。简而言之, 数字图书馆以现实资源的共享为目标[1]。数字图书馆的建设是当前图书馆建设的主要发展方向, 对人们的学习和生活有重要的现实意义。

实现图书自动分类是数字图书馆建设的重要一环。当前, 图书数量激增, 图书涉及到的领域知识越来越宽泛, 人工完成图书分类显得力不从心。因此, 将计算机自动化技术引入到图书分类领域, 实现图书自动分类, 已成为图书情报领域的研究热点, 能在很大程度上克服人力不足、相关人员专业知识薄弱等问题, 从而更高效准确地管理图书。

此外, 随着跨学科合作研究的不断增多和深入, 越来越多的跨领域成果涌现。与此相应, 越来越多的图书也不再局限于单个领域, 而是适用于分类法中的多个标签。若图书分类仍局限于单标签分类, 将导致图书被检索到的概率降低, 不利于图书的传播与共享。因此, 图书自动分类技术应充分考虑到多标签分类的情况, 更好地组织图书分类信息。

2 相关工作

文本分类的研究起源于20世纪50年代末, Luhn提出了词频的概念[2], 被视为是文本分类领域开创性的研究。总体而言, 国外文本分类的研究可以概括为以下4个发展阶段[3]: 第一阶段(1958年–1964年), 研究文本分类的可行性; 第二阶段(1965年–1974年), 对文本分类进行试验性研究; 第三阶段(1975 年–1989年), 对文本分类进行实用性研究; 第四阶段(1990 年至今), 面向互联网的文本分类研究。

通讯作者: 傅余洋子, ORCID: 0000-0003-1470-7720, E-mail: mg1414011@smail.nju.edu.cn。

*本文系国家自然科学基金项目"面向学术资源的TSD与TDC测度及分析研究"(项目编号: 71503121)和中央高校基本科研业务费重点项目"我国图书情报学科知识结构及演化动态研究"(项目编号: 20620140645)的研究成果之一。

在第四阶段之前，文本分类主要采用基于知识工程的方法，即由领域专家根据经验归纳出一系列的逻辑规则，以此作为计算机对文本进行分类时的依据。然而，这种方法的缺点很明显：分类的质量依赖于领域专家的水平，人力成本极高；规则不具备扩展能力，不同领域的规则需要不同领域的专家，且随着各领域的不断发展，规则需要实时更新。20 世纪 90 年代以来，一方面，随着信息时代的到来，依赖于人力的、基于知识工程的分类方法难以满足海量的、多样的文本信息；另一方面，人工智能技术快速发展，众多的学者将机器学习技术迁移到文本分类领域，文本分类开始向基于机器学习的分类系统转移。这类方法通过选择某些特征对文本进行形式化表示，设计并训练分类器对其进行分类，大大降低了人力成本，且具有更高的准确度和稳定性，因此逐渐成为文本分类领域的主流方法。目前，已有许多机器学习算法被应用到文本分类领域，如朴素贝叶斯分类法[4]、k-最近邻分类法[5]、支持向量机分类法[6]、神经网络分类法[7]等。

目前，文本分类技术取得了大量研究成果，也得到一定的应用。然而，其仍然面临着数据偏斜、非线性、多标签、标注瓶颈等问题[8]。其中多标签分类，指的是一个文本与不止一个类别相关联。在实际任务中，常常会出现多标签分类的情况。一般而言，对多标签分类问题的研究主要从以下三个角度出发[9]。

(1) 假设类别相互独立，在此前提下，最简单、最常用的方法为将多标签分类问题转换为多个二元分类问题，综合各二元分类的结果作为最终分类结果。如 Joachims 利用支持向量机算法实现了这种分类方法[10]。此外，还有基于排序的方法，在训练时学习得到一个排序函数，据此对文本和类别的匹配情况进行打分，将文本划分到分值高的类别。如 Crammer 等通过计算出每个类别的权重向量，进而计算文本特征向量与类别权重向量的内积，排序决定所属类别[11]。该类方法大多简单易行，且有大量高效算法可以直接利用，但对于类别间有关联的情况难以获得很好的性能。

(2) 考虑类别间的关联。一般通过构建主题模型解决多标签分类问题。如 Ueda 等提出一种产生式体系，包含任意两个类别间的关系[12]；Zhang 等提出双层主题分类模型，基于实例间的差异构建模型[13]。

(3) 利用半监督学习算法对未标记文本进行学习。这类方法综合考虑了同类别文本间的关系和不同类别文本间的关系。如 Liu 等通过求解带约束的非负矩阵获得最优的样本标签[14]。这类方法能有效地利用类别间的关系，但是学习过程较为复杂。总体而言，多标签分类问题比单标签分类问题更为复杂。

图书分类是文本分类的子领域，其理论基础与技术方法和文本分类相类似。然而，专门针对图书分类问题的研究相对较少，针对中文图书分类问题的研究则更少，且大多数处于试验阶段，尚未投入到实际应用之中。本文尝试将近年来发展迅速的深度学习算法引入中文图书自动分类领域，基于 LSTM 模型和字嵌入的方法构建分类系统，克服了手工分类中人力要求高、效率低、主观性强等问题，以及传统自动分类中预处理复杂、维护困难、可迁移性低等问题。

3　LSTM 模型介绍

长短时间记忆神经网络(Long Short Term Memory Neural Network, LSTM)最早是由 Hochreiter 等提出[15]，是一种基于时间序列的链式结构。Gers 等在原始模型的基础上加入了遗忘门[16]，这是 LSTM 模型的一个重要改进。近年来，Graves 对 LSTM 模型进一步的改良和推广[17]，从而使其进入蓬勃发展时期。本文使用的 LSTM 模型为当前业界公认的基本 LSTM 模型[18]。

LSTM 模型是循环神经网络(Recurrent Neural Network，RNN)的一种，针对 RNN 模型存在的梯度消失问题[19-20]而提出改进，用一个记忆单元替换原来 RNN 模型中的隐层节点。这个记忆单元的结构如图 1 所示[17-18]，由记忆细胞、遗忘门、输入门、输出门组成。记忆细胞负责存储历史信息，通过一个状态参数来记录和更新历史信息；三个门结构则通过 Sigmoid 函数决定信息的取舍，从而作用于记忆细胞。

具体而言，LSTM 模型主要涉及到以下计算过程，如公式(1)–公式(6)所示。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (4)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (5)$$

$$h_t = O_t \times \tanh(C_t) \quad (6)$$

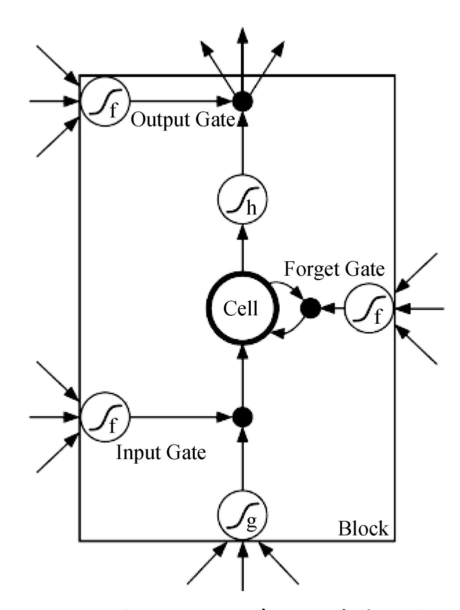


图 1　LSTM 模型的记忆单元的基本结构示意图

公式(1)–公式(3)分别是遗忘门、输入门、输出门的计算公式。在 t 时刻，门结构接受上一时刻记忆单元的输出 h_{t-1} 和当前时刻记忆单元的输入 x_t，与各自的权重矩阵相乘，然后加上偏置向量，通过 Sigmoid 函数产生一个 0 到 1 之间的值，对信息进行筛选。公式(4)–公式(5)是对 Cell 状态进行更新。公式(4)通过 tanh 函数对上一时刻记忆单元的输出 h_{t-1} 和当前时刻记忆单元的输入 x_t 进行计算，得出一个候选值，并由输入门决定将候选值的哪些信息更新到 Cell 状态中。同时，由遗忘门决定上一时刻 Cell 状态信息的保留情况，与更新的信息相加，得到当前时刻的 Cell 状态。公式(6)计算记忆单元最终的输出。通过 tanh 函数对当前时刻的 Cell 状态进行计算，使模型变为非线性的，并由输出门决定哪些信息将被最终输出。

可以看出，LSTM 模型采用累加的线性形式处理序列数据的信息，从而避免了梯度消失问题，也能学到长周期的信息[21]，克服了 RNN 模型的缺点。因此，LSTM 模型是处理时间序列数据常用的深度学习模型。

4　基于 LSTM 模型的中文图书分类系统设计

本文以 LSTM 模型为基础对中文图书分类系统进行设计，系统的整体架构如图 2 所示。系统共分为 5 个部分：输入层、Embedding 层、LSTM 隐层、Softmax 层、输出层。

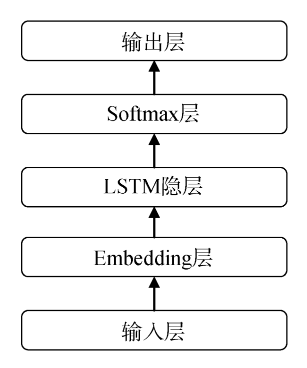


图 2　本文系统的整体架构示意图

整体架构中的第三部分即为本文重点采用的 LSTM 模型。在该部分，LSTM 隐层对之前部分处理得到的序列数据进行学习，利用 LSTM 模型的特性结合上下文信息，学习到的结果将传递给之后的部分，以进行分类预测。

除 LSTM 隐层外，第二部分 Embedding 层也是系统的重要组成部分。输入层将待分类文本统一为同等长度，并传递给 Embedding 层；Embedding 层采用字嵌入的方法，将传递来的文本中的每个字符转换为一个向量，从而将待分类文本转换为二维向量，传递给后续部分继续处理。字嵌入方法是在词嵌入方法的基础上提出的，而词嵌入方法则起源于 Hinton 提出的分布式表征的思想[22]。词嵌入将词表示为一个低维稠密向量，从而解决了维度灾难问题；并且可以通过余弦距离、欧式距离等方法计算词之间的相似度，从而克服了 One-hot 表示法这一类词表示方法无法反映词之间关系的问题。然而，完美的分词算法是不存在的[23]，故字嵌入的方法被提出。字嵌入将字符转换为低维稠密向量，继承了词嵌入的优点，同时避免了分词可能出现的问题。本文系统采用字嵌入的方法，使得整个系统是基于字符的，而不是基于特征词的，从而降低了系统的维护难度，不需要维护特征集合，也不需要在新特征加入后重新训练模型，而是可以采用增量学习的方法；此外，系统也可以方便地迁移到其他语言，提高了系统的应用价值。

Softmax 层用到了 Softmax 回归模型，对 LSTM 隐层传递来的信息进行学习，计算出待分类数据归属各类别的概率，传递给输出层，最终给出待分类文本的预测类别。Softmax 回归模型是 Logistic 回归模型的一般化形式，是常用的多分类算法。Softmax 回归模型拥

有很好的数学性质[24]；并且，其不仅能预测出对应的类别，还能计算出归属各类别的概率，方便更进一步的处理。该层也是损失函数的构建依据，系统整体采用 Adam 算法[25]进行优化。

5　实验与分析

实验数据为南京大学、同济大学、中国科学技术大学这三所高校的图书馆馆藏书目。这三所高校的图书馆书目检索系统均由江苏汇文软件有限公司[26]构建，具有基本一致的体系和格式，便于数据的获取和整理。使用 Python 语言[27]编写网络爬虫，分别从三所高校的图书馆书目检索系统爬取书目信息，并根据每条书目的机读格式(Machine Readable Catalog, MARC)获取特定字段的信息。主要抓取的字段及其含义如表 1 所示。

表 1　MARC 格式特定字段及含义

MARC 字段	含义
001	MARC 标识号
200	题名
330	摘要
606	主题词
690	中图分类号

针对《中国图书馆分类法》中的第一层分类进行实验。考虑到书目数量、实验规模等因素，没有将全部 22 个大类都纳入到实验中，而是选择 A(马克思列宁主义、毛泽东思想、邓小平理论)、C(社会科学总论)、F(经济)、N(自然科学总论)、X(环境科学、安全科学)这 5 个大类的书目为实验数据，涵盖了社会科学、自然科学等多个方面，具有一定的代表性。

本文从单标签分类、多标签分类两个角度进行实验。单标签分类指每个样本只属于一个特定的类别；而多标签分类指每个样本更倾向于属于多个类别[28]。利用单标签分类实验验证系统的可行性，再探索系统在多标签分类上的实用性。

实验环境为: CPU Intel Core i7-6700HQ，四核；内存 16GB；GPU NVIDIA GeForce GTX950M；显存 4GB；操作系统为 64 位 Ubuntu 16.04 LTS。

5.1　单标签图书分类实验

对单标签图书分类进行实验探索，即针对数据集中所有单标签条目进行分类实验。对数据进行合并、去重、去除多标签条目等操作，最终获得如表 2 所示的数据分布。根据类型抽样法将数据按 80%、10%、10%的比例划分为训练集、验证集、测试集，本节实验即在此数据集上进行。

表 2　单标签图书分类实验的数据分布

类标号	书目数
A	8 486
C	28 514
F	146 228
N	6 935
X	16 463
总计	206 626

考虑到不同的字段对书目内容的表达能力和涵盖情况不同，将探讨只选择题名字段、选择题名和主题词字段、选择题名和摘要字段、选择题名和主题词以及摘要字段这 4 种情况下的模型分类效果。选择基本单向 LSTM 模型；1 层 LSTM 隐层，每层隐层包含 128 个节点；每批处理的数据量为 128；训练过程采用早停原则，当模型在验证集上的损失值增大时则停止训练，且对整个训练集至多训练 1 000 轮，训练情况如图 3 所示。

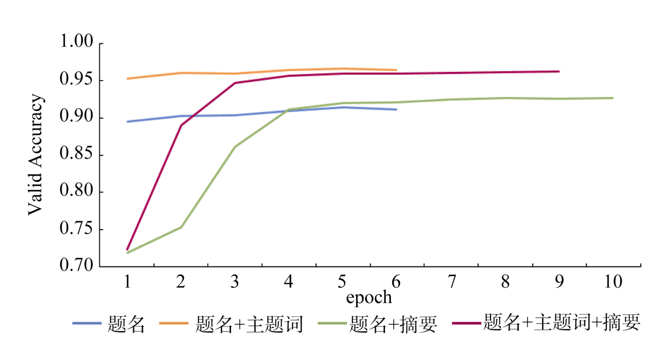


图 3　基于不同字段选择的模型在训练过程中在验证集上的准确率变化

图 3 反映了每训练完一遍整个训练集后的模型在验证集上的准确率变化趋势。观察只选择题名字段的模型与选择题名和主题词字段的模型，这两者在验证集上的准确率在训练初期便迅速趋于平稳。这是由于题名和主题词字段的字符数相对较少，模型学习到稳定状态的速度相对较快。从收敛情况来看，包含主题词字段的两个模型最终的准确率更高且相近，可见主题词字段对训练更优的模型有较大帮助。

当选择题名和主题词字段时，浅层的、单向的、基本 LSTM 模型在测试集上的分类准确率达到 97%左右，在具体类别上的 F1 值均高于 90%，处于较高水平。因此，综合实验情况、模型简单性、字段普遍性等因素，认为选择题名和主题词字段即可达到较好的分类效果，也具有可行性，后续的实验将基于这两个字段进行。

文献[29]同样对 A、C、F、N、X 等 5 大类进行实验，准确率为 95.94%，处于当前研究中的较高水平。本文则取得 96.97%的准确率，优于前人研究，证明本文系统确实具有可行性和应用价值。

5.2 多标签图书分类实验

对多标签图书分类进行实验探索，即针对数据集中所有条目进行分类实验，包括单标签条目和多标签条目。选择题名和主题词字段，对数据进行合并、去重的操作，最终获得如表 3 所示的数据分布。采用类型抽样法按 80%、20%的比例将数据集划分为训练集和测试集，本节实验即在此数据集上进行。

表 3 多标签图书分类实验的数据分布

类标号	书目数	类标号	书目数
A	8 101	A、X	5
C	25 595	C、F	1 217
F	133 401	C、N	69
N	6 461	C、X	50
X	15 642	F、N	49
A、C	38	F、X	684
A、F	111	N、X	21
A、N	4	C、F、X	3
总计			191 451

从表 3 可以看到，数据集中多标签的书目较少，只占数据总量的 1.18%。如果将多标签组合作为一个单独类别，采用单标签分类的方法进行分析。一方面，其对应的书目数量太少，难以学习到有效信息，或者存在过拟合风险；另一方面，只选择 5 个大类，若将全部类别都纳入考虑，则组合的数量将非常庞大，会给模型的训练带来难度。因此，将多标签组合作为一个单独类别是不具备可行性的，而因采取多标签分类的分析方法。对于《中国图书馆分类法》中同层次的类别而言，各类别之间是相对独立的，类别间没有什么关联。因此，本文将多标签分类问题转换为多个二元分类问题，即针对每个类别分别构建一个二元分类器，用于判断书目是否属于该类别。将训练集中所有属于该类别的数据标记为正类别，包括多标签的情况，而不属于该类别的数据标记为负类别，以此构建模型。

在单标签图书分类实验中，所有类的地位是一样的，即所有类的权重相同，在损失函数里的系数相同。而在本节实验中，一方面，大多数分类器的训练集存在类别不平衡的情况，正类别数据远远少于负类别；另一方面，实验目标在于尽量预测出所属的所有类别，故正类别的重要程度高于负类别，即正类别的误差成本应高于负类别。因此，笔者对类的权重进行调整，正类别与负类别的权重比为 15:1，损失函数中的对应系数据此进行调整。

笔者采用基本单向 LSTM 模型；2 层 LSTM 隐层，每层隐层包含 128 个节点；每批处理的数据量为 128；模型分别对各自的训练数据迭代学习 30 轮。训练情况如图 4 和图 5 所示。

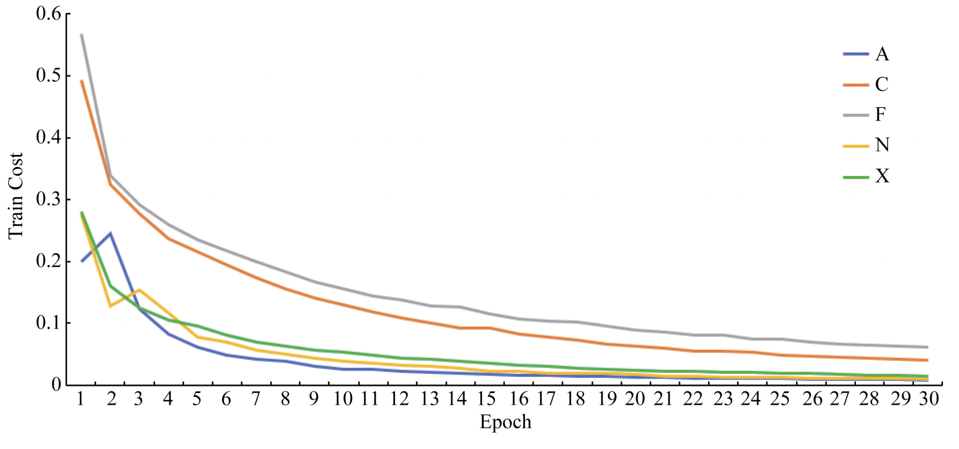


图 4 各二元分类器在训练集上的损失变化图

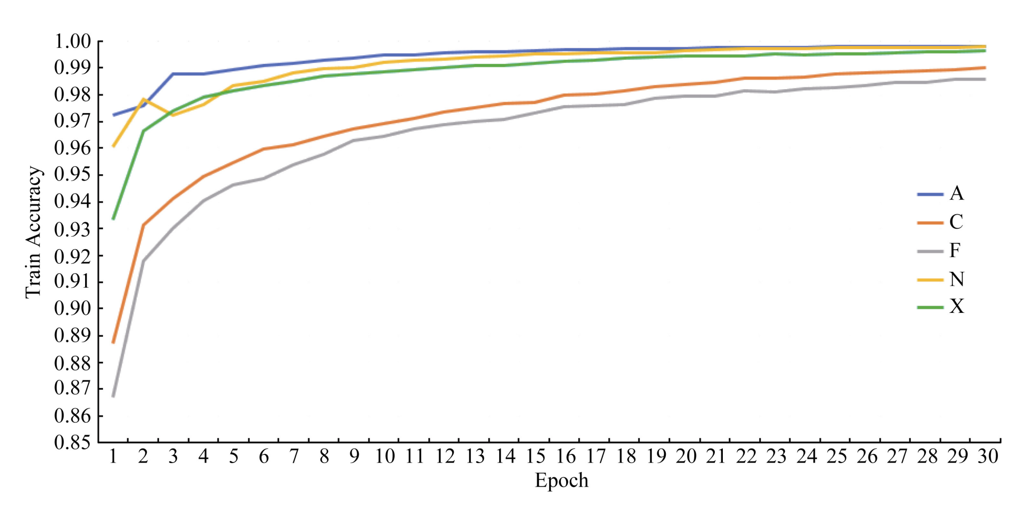


图 5　各二元分类器在训练集上的准确率变化图

可以发现两图反映出的信息是相符的，当训练到 15 轮左右时，各二元分类器在训练集上的损失和准确率都开始趋于平稳。结合各二元分类器对应的训练集来看，训练集包含的条目越少，其对应的二元分类器越快趋向于收敛，且停止训练时的损失值越低、准确率越高。由此可知，当训练集较大时，应适当增加训练的轮数以提升分类器的性能。

训练结束后，各二元分类器分别对测试集进行分类测试，并统计各类别的精度、召回率和 F1 值。统计过程中，包含所有属于某一类别的数据，包含单标签和多标签的情况，统计结果如表 4 所示。可以看到，15 个指标数据基本都在 85%以上，其中三分之二的指标数据在 90%以上。由此可见，各二元分类器的表现尚可，有一定的实际应用价值。

表 4　各类别的二元分类器在测试集上的测试情况表

类标号	精度	召回率	F1 值
A	91.23%	94.32%	92.75%
C	85.47%	93.61%	89.35%
F	95.85%	98.56%	97.19%
N	83.43%	90.17%	86.67%
X	88.88%	96.13%	92.36%

将各二元分类器在测试集上的指标数据与单标签分类实验进行对比，对比结果如图 6 所示。由于对类别权重进行调整，正类别的权重高于负类别，故对召回率的提升有促进作用。由图 6 可以发现，与单标签分类实验相比，召回率大多保持稳定或有所提升。但是，由于精度与召回率这两个指标相对矛盾，故精度显著下降。从 F1 值这一综合评价指标来看，分类表现在整体上略逊于单标签分类实验，但相差不大，仍处于较高水平。

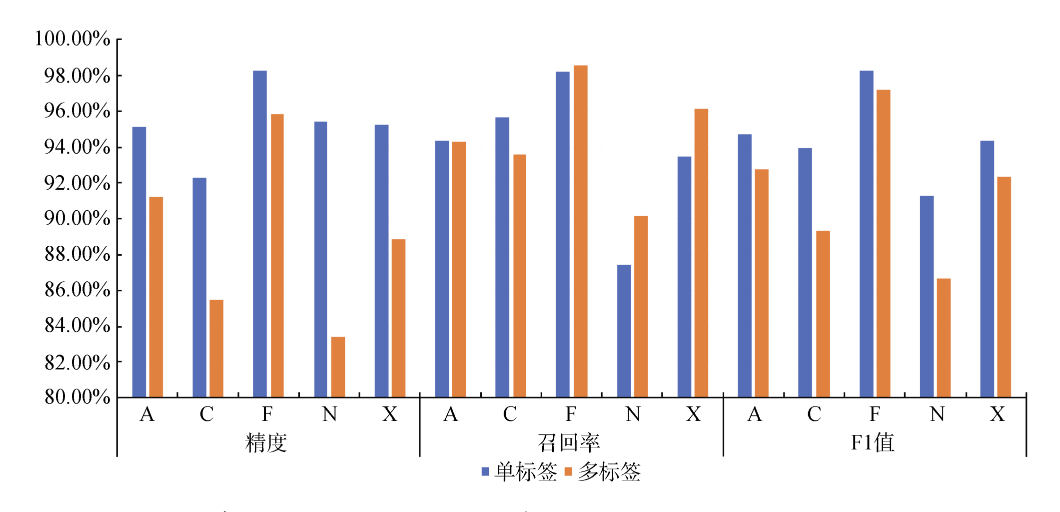


图 6　单标签实验与多标签实验在测试集各类别上的指标数据对比

针对测试集的分类结果，统计其中多标签条目的实际预测情况，统计结果如表 5 所示。

表 5 测试集中多标签条目的实际预测情况统计表

多标签项	实际存在数	预测情况		
		包含至少一个实际类别	包含全部实际类别	恰好等于实际类别
A、C	8	7	4	4
A、F	23	23	16	16
A、N	1	1	0	0
A、X	1	1	1	1
C、F	244	242	140	140
C、N	14	14	7	7
C、X	10	10	5	3
F、N	10	10	2	2
F、X	137	136	100	100
N、X	5	5	2	2
C、F、X	1	1	1	1
总计	454	450	278	276

可以发现，99.12%的多标签数据都至少预测出其中一个实际类别，61.23%的多标签数据预测出全部实际类别，60.79%的多标签数据恰好预测出了实际分类。同时，多标签项对应的实际数据越多，其被系统学习的情况越好，越有可能被全部预测出，反之则会导致偶尔性。由此可知，当增加更多多标签条目时，系统能获得更好的表现。此外，对整个测试集的预测情况进行统计分析，有 97.62%的数据至少被预测出一个实际分类，97.17%的数据的预测分类包含了全部实际分类，而 91.92%的数据被恰好完全预测正确。整体而言，系统在测试集上表现较好。

综上可知，对于多标签分类的中文图书分类任务，针对每个类别构建一个二元分类器，然后每个分类器采用浅层的、单向的、基于题名与主题词字段的、基本 LSTM 模型，这样的方法可以取得较好的分类表现，对单标签和多标签分类均有一定的实践意义。

6 结 语

本文将 LSTM 模型引入到中文图书分类问题中，与以往研究中的基于知识工程或基于传统机器学习的中文图书分类方法相比，本文系统具有如下优势。

(1) 预处理工作和后期维护工作简单。本文采用字嵌入方法，整个系统是基于字符序列构建的，故不需要分词、特征选择等过程。当出现新知识、新研究时，可以实现增量学习，无需重新训练模型。

(2) 充分利用 LSTM 模型的特点，递归神经网络适合处理序列数据。本文系统是对各字段直接连接而成的字符序列进行学习，一方面，避免了特征选择等过程中可能会误删重要信息等情况，对所有信息都进行学习；另一方面，相比于词袋模型未考虑词序等问题，LSTM 模型在处理序列数据时考虑了上下文信息，能更好地理解文本。

(3) 实验证明，无论是单标签图书分类，还是多标签图书分类，本文系统均有较好的表现。无论是整体的分类准确率，还是各类别的分类精度、召回率、F1 值，都达到了较高水平，有实际应用价值。

当然，本文研究也存在一些不足和改进空间，如测试环节未将全部类别纳入系统、分类粒度较粗等，这些将是笔者进一步的研究方向，以不断完善中文图书分类系统。

参考文献：

[1] 罗雪英. 也谈数字图书馆的建设目标[J]. 现代情报，2002, 22(12): 131-132. (Luo Xueying. Talking About the Construction Target of Digital Library [J]. Modern Information, 2002, 22(12): 131-132.)

[2] Luhn H P. Auto-encoding of Documents for Information Retrieval Systems[M]. IBM Research Center, 1958.

[3] 肖明. WWW科技信息资源自动标引的理论与实践研究[D]. 北京：中国科学院文献情报中心，2001. (Xiao Ming. Study on the Theory and Practice of Automatic Indexing of WWW Science and Technology Information Resources[D]. Beijing: National Science Library, Chinese Academy of Sciences, 2001.)

[4] Lewis D D, Ringuette M. A Comparison of Two Learning Algorithms for Text Categorization[C]//Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas. Information Science Research Institute, University of Nevada, 1994, 33: 81-93.

[5] Yang Y, Chute C G. An Example-based Mapping Method for Text Categorization and Retrieval[J]. ACM Transactions on Information Systems (TOIS), 1994, 12(3): 252-277.

[6] 陈立孚，周宁，李丹. 基于机器学习的自动文本分类模型研究[J]. 现代图书情报技术，2005(10): 23-27. (Chen Lifu,

Zhou Ning, Li Dan. Study on Machine Learning Based Automatic Text Categorization Model[J]. New Technology of Library and Information Service,2005(10): 23-27.)

[7] Weigend A S, Wiener E D, Pedersen J O. Exploiting Hierarchy in Text Categorization[J]. Information Retrieval, 1999, 1(3): 193-216.

[8] 苏金树，张博锋，徐昕．基于机器学习的文本分类技术研究进展[J]．软件学报，2006，17(9)：1848-1859. (Su Jinshu, Zhang Bofeng, Xu Xin. Advances in Machine Learning Based Text Categorization[J]. Journal of Software, 2006, 17(9): 1848-1859.)

[9] 吕小勇，石洪波．基于频繁项集的多标签文本分类算法[J]．计算机工程，2010，36(15)：83-85. (Lv Xiaoyong, Shi Hongbo. Multi-label Text Classification Algorithm Based on Frequent Item Sets[J]. Computer Engineering, 2010, 36(15): 83-85.)

[10] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features[A]// Machine Learning: ECML-98[M]. Springer, Berlin, Heidelberg, 1998: 137-142.

[11] Crammer K, Singer Y. A New Family of Online Algorithms for Category Ranking[C]// Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland. New York: ACM, 2002: 151-158.

[12] Ueda N, Saito K. Parametric Mixture Models for Multi-Labeled Text[A]//Advances in Neural Information Processing Systems[M]. MIT Press, 2003: 737-744.

[13] Zhang M, Zhou Z. Multi-Label Learning by Instance Differentiation[C]//Proceedings of the 22nd Conference on Artificial Intelligence. 2007: 669-674.

[14] Liu Y, Jin R, Yang L. Semi-supervised Multi-label Learning by Constrained Non-negative Matrix Factorization [C]// Proceedings of the 21st Conference on Artificial Intelligence, Boston, Massachusetts, USA. 2006, 6: 421-426.

[15] Hochreiter S, Schmidhuber J. Long Short-term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[16] Gers F A, Schmidhuber J, Cummins F. Learning to Forget: Continual Prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451-2471.

[17] Graves A. Supervised Sequence Labelling with Recurrent Neural Networks [D]. München: Technische Universität München, 2008.

[18] Zaremba W, Sutskever I, Vinyals O. Recurrent Neural Network Regularization [OL]. arXiv Preprint, arXiv: 1409.2329.

[19] Hochreiter S. Recurrent Neural Net Learning and Vanishing Gradient[J]. International Journal of Uncertainity, Fuzziness and Knowledge-Based Systems, 1998, 6(2): 107-116.

[20] Hochreiter S, Bengio Y, Frasconi P, et al. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-term Dependencies[A]// A Field Guide to Dynamical Recurrent Neural Networks[M]. Wiley-IEEE Press, 2001.

[21] 邱锡鹏．神经网络与深度学习[EB/OL]．[2017-04-21]. https://nndl.github.io/. (Qiu Xipeng. Neural Network and Deep Learning [EB/OL]. [2017-04-21]. https://nndl.github.io/.)

[22] Hinton G E. Learning Distributed Representations of Concepts[C]//Proceedings of the 8th Annual Conference of the Cognitive Science Society. 1986.

[23] Chung J, Cho K, Bengio Y. A Character-Level Decoder Without Explicit Segmentation for Neural Machine Translation[OL]. arXiv Preprint, arXiv:1603.06147.

[24] 周志华．机器学习[M]．北京：清华大学出版社，2016. (Zhou Zhihua. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.)

[25] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[OL]. arXiv Preprint, arXiv:1412.6980.

[26] HUIWEN Software [EB/OL]. [2017-02-13]. http://www.libsys.com.cn/.

[27] Python Software Foundation [EB/OL]. [2017-02-12]. https://www.python.org/.

[28] 李思男，李宁，李战怀．多标签数据挖掘技术：研究综述[J]．计算机科学，2013, 40(4): 14-21. (Li Sinan, Li Ning, Li Zhanhuai. Multi-label Data Mining: A Survey[J]. Computer Science, 2013, 40(4): 14-21.)

[29] 王昊，严明，苏新宁．基于机器学习的中文书目自动分类研究[J]．中国图书馆学报，2010,36(6)：28-39. (Wang Hao, Yan Ming, Su Xinning. Research on Automatic Classification for Chinese Bibliography Based on Machine Learning[J]. Journal of the Library Science in China, 2010, 36(6): 28-39.)

作者贡献声明：

邓三鸿，傅余洋子，王昊：提出研究思路，设计研究方案；
傅余洋子：设计模型，进行实验，分析数据，起草论文；
邓三鸿：论文最终版本修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: mg1414011@smail.nju.edu.cn。

[1] 傅余洋子. 单标签分类实验数据.zip. 单标签图书分类实验数据集.

[2] 傅余洋子. 多标签分类实验数据.zip. 多标签图书分类实验数据集.

收稿日期: 2017-05-27

收修改稿日期: 2017-07-03

Multi-Label Classification of Chinese Books with LSTM Model

Deng Sanhong　Fu Yuyangzi　Wang Hao

(School of Information Management, Nanjing University, Nanjing 210023)

(Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing 210023, China)

Abstract: [**Objective**] This paper proposes a new method to automatically cataloguing Chinese books based on LSTM model, aiming to solve the issues facing single or multi-label classification. [**Methods**] First, we introduced deep learning algorithms to construct a new classification system with character embedding technique. Then, we trained the LSTM model with strings consisting of titles and keywords. Finally, we constructed multiple binary classifiers, which were examined with bibliographic data from three universities. [**Results**] The proposed model performed well and had practical value. [**Limitations**] We only analyzed five categories of Chinese bibliographies, and the granularity of classification was coarse. [**Conclusions**] The proposed Chinese book classification system based on LSTM model could preprocess data and learn incrementally, which could be transferred to other fields.

Keywords: LSTM Model　Deep Learning　Character Embedding　Book Automatic Classification　Multi-label Classification